



Murphy, J. L., Puttick, M. N., O'Reilly, J. E., Pisani, D., & Donoghue, P. C. J. (2021). Empirical distributions of homoplasy in morphological data. *Palaeontology*, 64(4), 505-518.
<https://doi.org/10.1111/pala.12535>

Publisher's PDF, also known as Version of record

License (if available):
CC BY

Link to published version (if available):
[10.1111/pala.12535](https://doi.org/10.1111/pala.12535)

[Link to publication record in Explore Bristol Research](#)
PDF-document

This is the final published version of the article (version of record). It first appeared online via Wiley at <https://doi.org/10.1111/pala.12535>. Please refer to any applicable terms of use of the publisher.

University of Bristol - Explore Bristol Research

General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available:
<http://www.bristol.ac.uk/red/research-policy/pure/user-guides/ebr-terms/>

EMPIRICAL DISTRIBUTIONS OF HOMOPLASY IN MORPHOLOGICAL DATA

by JODIE L. MURPHY^{1,*}, MARK N. PUTTICK^{1,2,*} , JOSEPH E. O'REILLY^{1,3} ,
 DAVIDE PISANI^{1,4}  and PHILIP C. J. DONOGHUE¹ 

¹School of Earth Sciences, University of Bristol, Life Sciences Building, Tyndall Avenue, Bristol BS8 1TH, UK; phil.donoghue@bristol.ac.uk

²School of Biological Sciences, University of Bristol, Life Sciences Building, Tyndall Avenue, Bristol BS8 1TH, UK

³MRC Institute of Genetics & Molecular Medicine, University of Edinburgh, Edinburgh, UK

⁴Department of Biochemistry & Biological Sciences, University of Bath, Bath, UK

*These authors contributed equally to this study

Typescript received 8 June 2020; accepted in revised form 4 January 2021

Abstract: Cladistic datasets of morphological characters are comprised of observations that exhibit varying degrees of consistency with underlying phylogenetic hypotheses, reflecting the acquisition and retention of character states (highly consistent characters), or the convergent evolution and loss of character states (less consistent characters). The consistency between phylogenetic history and individual character histories has a bearing both on the evolutionary process and on the relative ease with which phylogenetic history may be inferred from morphological data. We surveyed 486 tetrapod morphological cladistic datasets to establish an empirical distribution of consistency among characters and datasets. Average dataset size has increased in the number of characters and taxa through time. The Consistency Index measure of homoplasy decreases as more characters are added but the most significant decreases result from the addition of taxa. Retention

Index and Homoplasy Excess Ratio remain relatively constant with changes in taxa and character number. Our sampling of larger datasets confirms that the positive relationship between dataset size and homoplasy is primarily caused by an increase in taxa, not an increase in characters. Genealogies of cladistic data matrices for early vertebrates, scolidophorans and crocodilians, which have been modified in succession, show a trend of generally consistent quality through research time. Thus, we find no support for the widely shared conjecture that in the search for phylogenetic resolution, high quality phylogenetic characters are quickly exhausted, with subsequent research leading to the inclusion of potentially misleading characters exhibiting high levels of homoplasy.

Key words: homoplasy, Consistency Index, Retention Index, morphology, categorical, cladistic.

THE reconstruction of evolutionary relationships among species is based on the premise that their shared characters reflect shared ancestry, and that similarity is known as homology. However, taxa can share superficially similar characters that are derived from separate ancestors and homologous characters can be absent as a consequence of loss, both of which contribute to the phenomenon of homoplasy which can misguide phylogenetic inference. When the proportion of homoplastic characters in a dataset is large, the homologous phylogenetic signal can be difficult to discriminate from homoplasy, resulting in recovery of an incorrect topology (Scotland & Steel 2015) that may, nevertheless, exhibit spuriously high statistical support in measures of phylogenetic fidelity, such as both the bootstrap or posterior probability (Philippe *et al.* 2011).

Cladistic matrices tend to grow in size over research time, but it is unclear whether newly added characters confer more phylogenetic signal than noise. Previous surveys have demonstrated that the proportion of homoplasy

within a dataset increases with the number of characters and taxa included (Sanderson & Donoghue 1989; Archie 1996; Hauser & Boyajian 1997), interpreted by some to reflect the rapid exhaustion of highly informative, low homoplasy, phylogenetic characters in the compilation of phylogenetic datasets and, perhaps, the increasing desperation of phylogeneticists seeking characters to resolve otherwise poorly supported clades (Scotland *et al.* 2003; Nelson 2004; Scotland & Steel 2015).

Homoplasy is not merely a nuisance parameter that plagues the efforts of phylogeneticists, but an interesting evolutionary phenomenon worthy of study in its own right (Hauser & Boyajian 1997; Sanderson & Donoghue 1989, 1996). Recently, the nature and distribution of homoplasy, as well as the efficacy of its measures, have come into sharp focus as a measure of the phylogenetic informativeness of cladistic matrices as researchers have attempted to benchmark the performance of competing methods for the phylogenetic analysis of morphological

data (O'Reilly *et al.* 2016, 2018a, b; Brown *et al.* 2017; Puttick *et al.* 2017a, b, 2019; Goloboff *et al.* 2018a, b; Sansom *et al.* 2018; Keating *et al.* 2020). Much of this debate has been conducted within the context of experiments based on simulated datasets, a conventional approach in molecular phylogenetics where empirically realistic data can be simulated based on well-founded theoretical expectations and empirical observations of sequence evolution. In contrast, the evolution of morphology lacks a coherent statistical description that can be modelled. Thus, the level of homoplasy observed in real datasets has been invoked to assay the empirical realism of simulated morphological data. Unfortunately, the most recent detailed surveys of homoplasy in morphological datasets were conducted more than 20 years ago (Sanderson & Donoghue 1989, 1996; Hauser & Boyajian 1997), at a time when datasets were limited in size by computing power and efficiency. Indeed, both the number and size of datasets investigated in these surveys were small by contemporary standards where matrices comprising hundreds of taxa and characters, are not unusual (O'Reilly *et al.* 2018a). Furthermore, early studies measured homoplasy as a matrix average, rather than as a proportional distribution which might provide a more precise characterization of empirical data (Goloboff *et al.* 2018a, b; O'Reilly *et al.* 2018a). Finally, previous surveys have focused on the Consistency Index (CI) as a measure of homoplasy, while this is just one of several widely used metrics, including the Homoplasy Excess Ratio (HER) (Archie 1989) and the Retention Index (RI) (Kluge & Farris 1969; Farris 1989) which some suggest to be a more accurate characterization of the homoplasy in empirical morphological data (Sansom & Wills 2017).

We undertook a survey of homoplasy in empirical morphological datasets, encompassing the greater number and scale of contemporary studies, using the CI, HER and RI. Our results provide a better understanding of morphological homoplasy and a better characterization of this phenomenon as a target for simulation studies.

MATERIAL AND METHOD

We based our study on two published databases of morphological cladistic datasets (Wright *et al.* 2016; Sansom *et al.* 2018), for a total of 486 datasets, excluding a small number of MRP (matrix representation with parsimony) supertree matrices and categorical molecular datasets. These datasets derive ultimately from a database curated by Graeme Lloyd (<http://www.graemetlloyd.com/matr.html>). The year of publication and the dimensions of these datasets, in terms of numbers of taxa, characters and character states, are summarized in Figure 1. Assumptions were removed from all datasets (e.g. differential character

weighting and character state ordering) to eliminate the effects of subjective opinions on the true history of characters or their relative value as phylogenetic markers. All the other parsimony analyses were implemented in PAUP4a build 166 for Unix/Linux (Swofford 2002). We tested whether the retained datasets conveyed significant clustering information using permutation tail probability (PTP) tests. For the PTP test, 20 000 permutations with heuristic search were used, with one random addition replicate for each search, retaining just one most-parsimonious tree for each permuted dataset.

For all analyses, a most parsimonious tree was inferred for each dataset using a heuristic search, with 10 000 random addition sequences and swapping trees using TBR in PAUP. The *multree* option was turned off to speed up searches. This implies that for each island of multiple most parsimonious trees, only one tree was retained. This strategy is adequate for our study as all trees in an island of most-parsimonious trees will have the same tree length, CI, RI and HER.

Constant and parsimony-uninformative sites were removed prior to analysis, as they can inflate CI and RI values (Sanderson & Donoghue 1989). The reuse of matrices can also bias the calculation of CI and RI values and so the independence of matrices was determined following the approach of Sansom *et al.* (2018), where matrices with a high proportion of taxa shared with one or more other matrices, are removed iteratively until the only remaining matrices have at least 50% unique taxa and on average all matrices had 75% unique taxa; this left 364 independent matrices.

Measures of homoplasy

We measured homoplasy based on the CI, RI and HER which are calculated as follows:

$$CI = \frac{m}{s}$$

$$RI = \frac{g - s}{g - m}$$

where m is the minimum number of possible changes, s is the observed number of changes, and g is the maximum number of changes for a tree computed using maximum parsimony (Kluge & Farris 1969; Farris 1989). Both indices are scaled from zero to one, where a value of one reflects no homoplasy and the proportion of homoplasy increases as the value gets smaller (i.e. homoplasy increases as the index value decreases). CI and RI can be used to characterize the entire matrix or each character individually. CI is a blunt measure of homoplasy, in binary characters reducing from 1 to $\frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \dots$

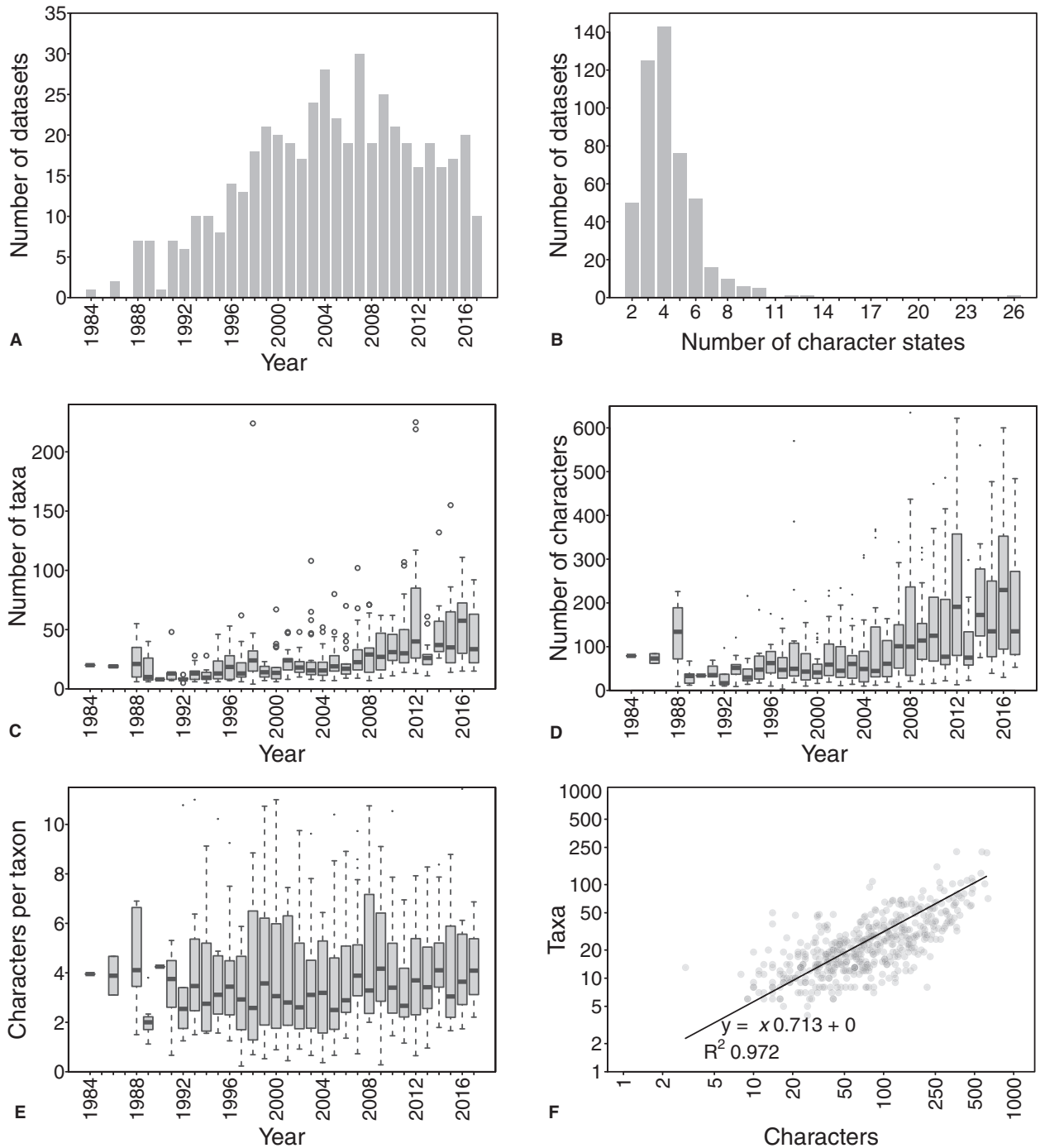


FIG. 1. Characteristics of the datasets analysed in this study. The datasets were compiled by previous meta-analyses (Wright *et al.* 2016; Sansom *et al.* 2018) but filtered to exclude a small number of MRP supertree matrices and a categorical molecular dataset. A, number of datasets relative to their year of publication. B, number of datasets relative to the number of character states used. C, number of taxa within datasets relative to their year of publication. D, number of characters within datasets relative to their year of publication. E, number of characters per taxon within datasets relative to their year of publication. F, number of taxa and characters within datasets.

(reflecting increasing homoplasy) with every additional state change beyond the minimum possible number. Thus, characters can have a low CI even if they have a

high consistency in one part of the tree but not another. RI reflects the congruence of a character with a topology in a way that better reflects the fact that homoplastic

characters can nevertheless be phylogenetically informative. Hence, a phylogenetically informative character with a low CI might still have a high RI. In this sense RI is a less blunt measurement of homoplasy.

We calculated HER using the following formula:

$$\text{HER} = 1 - \frac{l - l_{\min}}{\text{MeanNS} - l_{\min}}$$

The HER expresses the excess homoplasy in a dataset with reference to the level of homoplasy we expect to observe in random datasets of equivalent size. In the formula above, l represents the parsimony length of the optimal tree; l_{\min} is the theoretical minimal tree length for a dataset with that tree, number of taxa, characters and character states; and MeanNS (mean number of steps) is the average tree length of a set of trees generated from a set of randomizations of the considered dataset. Expressing excess homoplasy with reference to the homoplasy of randomized datasets is an alternative way to explore homoplasy in datasets. HER values, in the largest majority of cases, are expected to be bound between 0 and 1 (with 1 indicating datasets with no homoplasy) but the HER may produce negative values (Farris 1991). Archie (1989) observed that this could occur due to sampling error, and our analyses (below) show that, indeed, negative HER values are associated with datasets that do not convey clustering signal (i.e. random at the PTP test). To estimate HER, we used custom Perl scripts to first use PAUP to calculate, for each dataset, a most parsimonious tree and its length (l), the theoretical minimal tree length (l_{\min}) and to run a PTP test. After that, a second Perl script used the output from the PTP test to estimate MeanNS, and to calculate HER using the formula above. The PTP test was run as described above.

We analysed patterns of CI, RI and HER in R (R Core Team 2016). As CI and RI are bounded between 0 and 1, the relationship of CI and RI against the number of characters and taxa was analysed using beta regressions using the *betareg()* function in the *betareg* R package (Cribari-Neto & Zeileis 2010). We also analysed HER with beta regressions and converted negative values to zero prior to analyses. To analyse the within-matrix homoplasy distributions, the per character CI and RI values for each matrix were separated into one of 10 bins between 0 and 1, with the bins increasing in increments of 0.1.

To test the hypothesis that character consistency decreases over sampling time we identified three genealogies of matrices that are linked by successive modification based necessarily on our familiarity with the originating studies; this required extending our analyses beyond tetrapods. These matrices were developed to resolve the relationships among early vertebrates (Forey 1995; Janvier

1996; Donoghue *et al.* 2000; Shu *et al.* 2003; Gess *et al.* 2006; Sansom *et al.* 2010; Conway Morris & Caron 2014; Gabbott *et al.* 2016), scalidophorans (Wills 1998; Dong *et al.* 2004, 2005; Harvey *et al.* 2010; Wills *et al.* 2012; Ma *et al.* 2014; Shao *et al.* 2018) and crocodilians (Brochu 1997, 1999, 2006, 2011; Gold *et al.* 2014; Lee & Yates 2018). We reduced each of these to a common set of taxa to control for the effects of taxa on character consistency, and inapplicable and polymorphic character states were converted to a missing state. Otherwise, the same protocol described above was applied to characterizing homoplasy in these datasets.

The R and Perl scripts used, metadata on the datasets analysed and the datasets used in the evaluation of matrix genealogies, are available in Murphy *et al.* (2021).

RESULTS

The relationship between the year of publication and the number of taxa and characters comprising a dataset clearly demonstrates that average number of characters and taxa has increased over time (Fig. 2A, B); each year the size of the average matrix has increased by $1.06\times$ parsimony-informative characters and $1.05\times$ taxa. This growth in matrices through time is mainly the result of the augmentation of existing matrices as unique matrices have no significant increase in the total number of characters or taxa through time. Linear regression analysis of the relationship between the number of characters and number of taxa in a matrix identified a 2.03:1 ratio of characters to taxa (Fig. 2C), with a ratio of 2.07:1 when parsimony uninformative characters are included.

CI is negatively impacted by the inclusion of more taxa (Fig. 3A) and more characters (Fig. 3B) in a dataset, reflecting increased homoplasy, while RI (Fig. 3C, D) and HER (Fig. 3E, F) values remain relatively consistent. However, RI values slightly decrease (Fig. 3C, D) and HER increases slightly (Fig. 3E, F) as more taxa and characters are added. Beta regressions identified a highly significant negative association ($p < 0.001$) for each of the relationships between matrix-wide CI and RI and the number of taxa and characters (Fig. 3) in a dataset. However, the amount of variance explained by the number of characters or taxa, shown by R^2 , and the amount the slope deviates from zero to negative values, are highly variable (Table 1). In contrast, much less of the variance observed in RI can be explained by the number of taxa or characters (R^2 values of 0.02–0.05). HER increases with higher numbers of characters and taxa but neither is a good predictor of HER values in a data matrix ($R^2 = <0.03$). These models can predict the expected CI, RI and HER for a matrix given the number of taxa or characters (Table 2).

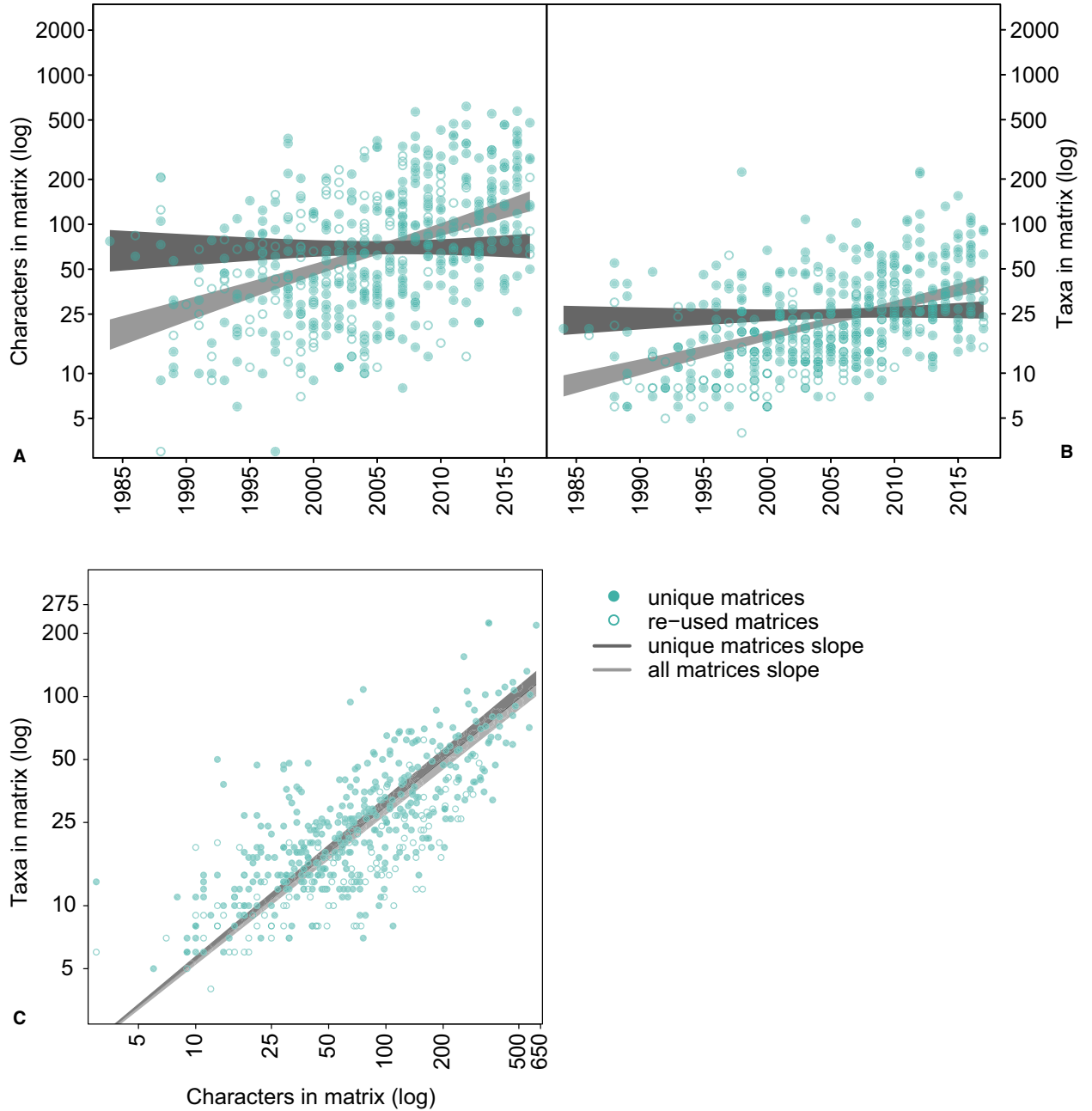


FIG. 2. Trends in the number of characters and taxa in matrices through time. The plots include information for all matrices and unique matrices are those that we consider to be independent. When all matrices are considered, the beta regression model estimates show an increasing number of characters and taxa through time. However, this may be an artefact of expansion of published matrices as unique matrices do show a flat relationship of character and taxa number through time. A, numbers of characters in a matrix plotted against year of publication. B, numbers of taxa in a matrix plotted against year of publication. C, numbers of taxa in a matrix plotted against numbers of characters in a matrix.

When the number of taxa and characters are considered together in a multivariate regression, the number of taxa is a significant negative covariate of CI but the slope for character number does not significantly differ from zero (Table 1). The number of taxa is not a significant

correlate of RI in the multivariate regression with unique character matrices, and character number is only a significant negative correlate when all matrices (unique and reused) are included. There was no significant collinearity between the number of taxa and characters in the

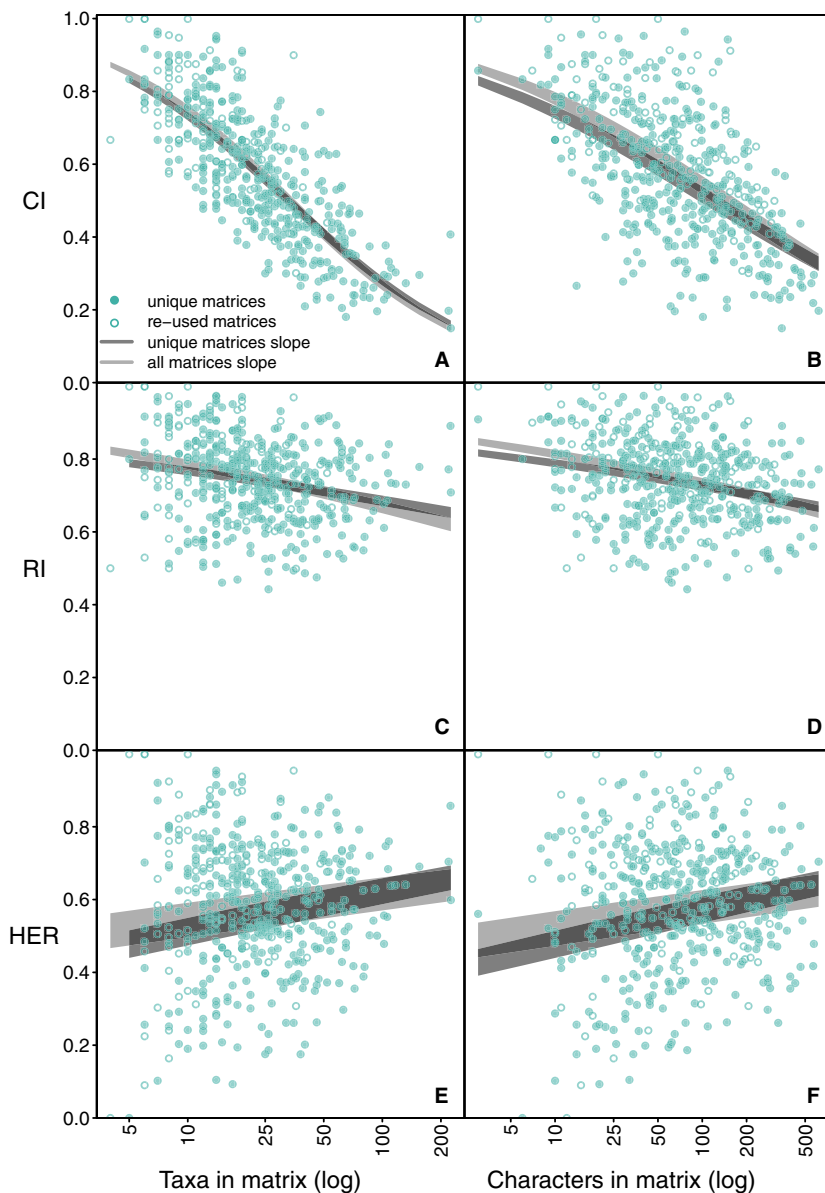


FIG. 3. Relationships of whole-matrix CI (A–B), RI (C–D) and HER (E–F) values with taxa and character number. CI and RI decrease as matrices grow in size, but HER increases positively with matrix size.

multiple beta regression with the generalized variance inflation factors (range 2.16–2.45) all below 5 (James *et al.* 2013).

A bimodal distribution was recovered for the per character CI distribution (Fig. 4). The first peak, at the 0.9–1 bin, accounts for an average of *c.* 40% of the characters within datasets and the second peak, within the CI 0.4–0.5 bin, accounts for *c.* 25% of the characters. Less than 5% of the characters in datasets have CI 0.5–0.9, while characters with CI ≤ 0.4 gradually decline proportionally as overall matrix CI decreases. The per character RI distribution is unimodal with a peak of *c.* 40% of characters at RI 0.9–1 (Fig. 3). For RI 0.3–0.9, an approximately consistent proportion of characters (*c.* 5–10%) is observed in each bin. Characters with the most

homoplasy, RI bin 0–0.1, account for *c.* 7% of the characters in a dataset.

All empirical estimates of CI, RI and HER for the matrix genealogies fall within the range of values expected based on our regression analyses of the broader dataset of matrices. As taxa number, not character size, is the significant correlate of CI and RI, these remain constant as characters are added to matrices. The early vertebrate matrices show the worst RI and HER scores relative to the expectation. The biggest change in CI, RI and HER from one matrix to the next are the decreases in each value from Gold *et al.* (2014) to Lee & Yates (2018) crocodilian matrices.

There are only small changes in matrix-wide values of homoplasy for the genealogies of matrices that we

TABLE 1. Summary of beta regression for CI, RI and HER against the number of characters and taxa using univariate and multiple regressions (separate results are summarized for all matrices and those with unique characters).

	Estimate	SE	Z	<i>p</i>	Precision (Phi)	Pseudo R ²
CI ~ Taxa						
Intercept	3.14	0.12	26.34	<0.001	13.36	0.48
Slope	−0.91	0.04	−24.15	<0.001		
CI ~ Taxa (unique matrices)						
Intercept	2.96	0.13	22.05	<0.001	15.64	0.56
Slope	−0.85	0.04	−20.97	<0.001		
CI ~ Characters						
Intercept	2.37	0.14	17.28	<0.001	8.57	0.30
Slope	−0.48	0.03	−15.21	<0.001		
CI ~ Characters (unique matrices)						
Intercept	2.05	0.15	13.93	<0.001	9.87	0.32
Slope	−0.43	0.03	−12.87	<0.001		
RI ~ Taxa						
Intercept	1.90	0.13	14.83	<0.001	11.89	0.07
Slope	−0.26	0.04	−6.57	<0.001		
RI ~ Taxa (unique matrices)						
Intercept	1.61	0.14	11.26	<0.001	14.0	0.04
Slope	−0.18	0.04	−4.22	<0.001		
RI ~ Characters						
Intercept	1.94	0.13	15.49	<0.001	12.07	0.08
Slope	−0.20	0.03	−7.07	<0.001		
RI ~ Characters (unique matrices)						
Intercept	1.65	0.13	12.43	<0.001	14.23	0.06
Slope	−0.15	0.03	−4.93	<0.001		
HER ~ Taxa						
Intercept	−0.12	0.18	−0.68	0.497	4.22	0.005
Slope	0.13	0.06	2.31	0.021		
HER ~ Taxa (unique matrices)						
Intercept	−0.41	0.19	−2.12	0.034	5.60	0.02
Slope	0.20	0.06	3.38	<0.001		
HER ~ Characters						
Intercept	−0.16	0.17	−0.93	0.352	4.23	0.007
Slope	0.10	0.04	2.63	0.009		
HER ~ Characters (unique matrices)						
Intercept	−0.47	0.18	−2.67	0.008	5.67	0.02
Slope	0.17	0.04	4.07	<0.001		
CI ~ Taxa + Characters						
Intercept	3.24	0.13	25.49	<0.001	13.40	0.48
Slope (taxa)	−0.87	0.05	0.05	<0.001		
Slope (characters)	−0.04	0.04	−1.14	0.254		
CI ~ Taxa + Characters (unique matrices)						
Intercept	2.97	0.14	21.50	<0.001	15.64	0.56
Slope (taxa)	−0.84	0.06	−14.08	<0.001		
Slope (characters)	−0.013	0.04	−0.33	0.74		
RI ~ Taxa + Characters						
Intercept	2.03	0.13	15.15	<0.001	12.17	0.09
Slope (taxa)	−0.12	0.06	−2.01	0.04		
Slope (characters)	−0.14	0.04	−3.28	0.001		
RI ~ Taxa + Characters (unique matrices)						
Intercept	1.71	0.15	11.67	<0.001	14.27	0.06
Slope (taxa)	−0.06	0.06	−0.91	0.364		
Slope (characters)	−0.12	0.05	−2.57	0.01		

(continued)

TABLE 1. (Continued)

	Estimate	SE	Z	<i>p</i>	Precision (Phi)	Pseudo R ²
HER ~ Taxa + Characters						
Intercept	−0.20	0.18	−1.08	0.28	4.24	0.007
Slope (taxa)	0.05	0.08	0.56	0.57		
Slope (characters)	0.08	0.06	1.36	0.18		
HER ~ Taxa + Characters (unique matrices)						
Intercept	−0.52	0.20	−2.64	0.008	5.68	0.02
Slope (taxa)	0.05	0.09	0.57	0.57		
Slope (characters)	0.14	0.06	2.27	0.02		

TABLE 2. Predicted CI, RI and HER given a number of taxa or characters in a matrix estimated from univariate beta regressions applied to the empirical data.

	10	25	50	100	250	500
Number of taxa						
Predicted CI	0.75	0.268	0.137	0.08	0.043	0.023
Predicted CI (unique matrices)	0.73	0.279	0.151	0.09	0.052	0.029
Predicted RI	0.786	0.668	0.614	0.57	0.526	0.481
Predicted RI (unique matrices)	0.767	0.685	0.649	0.62	0.59	0.559
Predicted HER	0.544	0.616	0.643	0.663	0.683	0.701
Predicted HER (unique matrices)	0.512	0.623	0.664	0.694	0.722	0.749
Number of characters						
Predicted CI	0.78	0.542	0.433	0.354	0.282	0.22
Predicted CI (unique matrices)	0.743	0.516	0.418	0.347	0.283	0.226
Predicted RI	0.814	0.733	0.695	0.665	0.633	0.6
Predicted RI (unique matrices)	0.789	0.726	0.698	0.676	0.653	0.629
Predicted HER	0.52	0.58	0.603	0.62	0.636	0.653
Predicted HER (unique matrices)	0.477	0.572	0.609	0.636	0.662	0.688

identified (Table 3; Fig. 5). The total range of matrix-wide CI values of matrix genealogies is lower than 0.1 for early vertebrates (0.069), scalidophorans (0.068) and crocodilians (0.085). Similar trends are seen with RI values, but variance is higher for HER through time, even though the patterns of increment and decrement are comparable across all three measures.

There is a significant decrease in the proportion of parsimony-informative characters across the three matrix genealogies. This is likely to be an artefact of pruning matrices with increasing numbers of taxa, to the same small core set of taxa. Nevertheless, the percentage of parsimony informative characters decreases each year by 16% (early vertebrates), 5% (scalidophorans) and 1% (crocodilians). Otherwise, there is no significant change in matrix-wide CI values across the matrix genealogies. For the crocodilian matrices, there is a significant trend of decreasing CI value ($p = 0.02$) particularly associated with the most recent matrix generation within the genealogy (Lee & Yates 2018), but there is no significant trend in matrix-wide CI across the genealogy of crocodilian matrices. All genealogies of datasets show a trend of

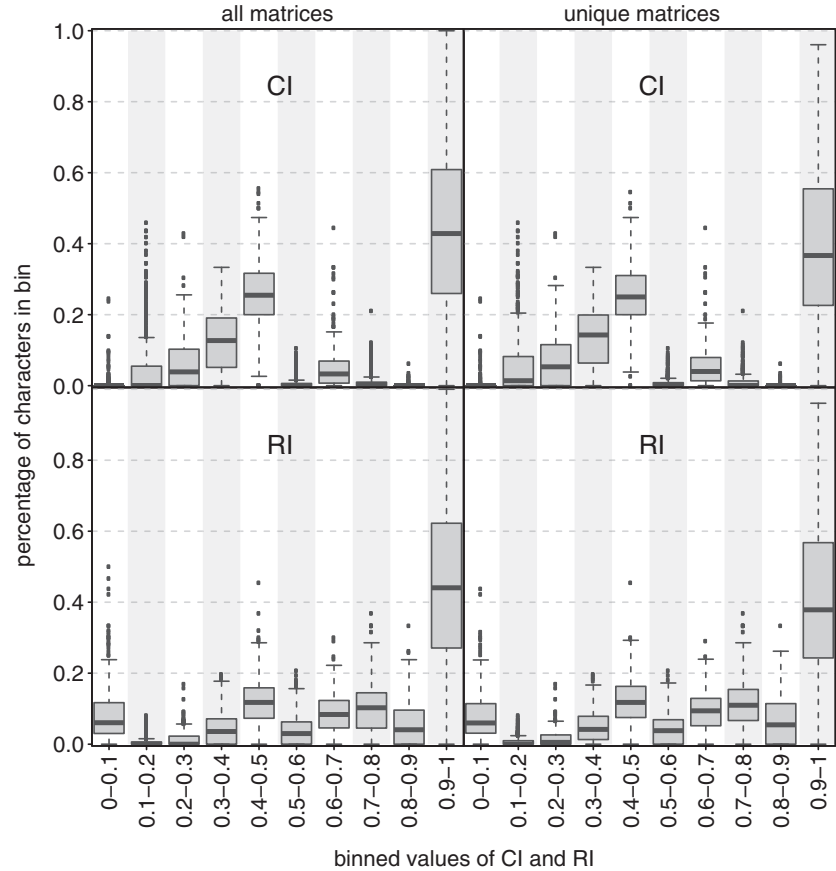
decreasing matrix-wide RI value through time, but this is only significant for the early vertebrate matrices. All dataset genealogies show a decreasing trend of HER values from time, but none of these is significant.

DISCUSSION

Homoplasy and matrix size

As observed by previous metaanalyses (Kluge & Farris 1969; Archie 1989; Sanderson & Donoghue 1989, 1996; Klassen *et al.* 1991; Archie & Felsenstein 1993; Lamboy 1994; Givnish & Sytsma 1997; Hauser & Boyajian 1997; Wiens 2004) datasets that have more taxa and/or more characters contain more homoplasy, reflected in decreasing CI and RI (Fig. 3A–D). Increasing taxon numbers in a matrix has a greater negative effect on CI values compared to increasing character number (Table 1). A significant negative relationship was also identified between larger datasets and the overall RI values (Fig. 3C, D). However, as the number of characters and taxa increase,

FIG. 4. Distribution of per-character CI and RI values between 0 and 1, pooled across all matrices and placed in 0.1 size bins.



there is a greater decrease in CI compared to RI. This suggests that the RI value is not greatly affected by the size of the dataset (Wiens 2004) while CI is greatly affected. In particular, the number of characters and taxa explain around 45% of the variance in CI values compared to only 5% for RI. It is likely that the discrepancy between the effect of matrix size on CI and that on RI or HER, reflects the difference in how these indexes are calculated and what they actually measure. For HER, the excess of homoplasy is measured with reference to that observed in random data.

Given the approximately flat trend in overall RI and HER (but with different directionalities; Fig. 3C–F) there is no evidence to suggest that as more data are added to morphological datasets their phylogenetic signal decreases (Wiens 2004). While in individual cases this might have occurred, for example in the dataset of Lee & Yates (2018) (Fig. 5C), this is far from representing a trend. Instead, the general observed trend seems to suggest that while homoplasy increases through generations of dataset development, this is matched broadly with an increase in phylogenetic signal. Thus, as datasets become larger, their excess homoplasy, with reference to that of random datasets, tends to decrease (Fig. 3E, F). Furthermore, there is a significant relationship between the number of

characters and the number of taxa, with a character to taxon ratio of 2.11:1 (Fig. 2).

We suggest that the difference in the behaviour of the RI and HER in response to increasing numbers of taxa and characters, could be explained with reference to the fact that HER measures the homoplasy in a dataset with reference to that observed in random data. As the number of taxa and characters increase, the trend observed in RI tells us that only a small amount of homoplasy is added to the data (Fig. 3C, D), while the HER trend indicates that despite some addition of homoplasy (RI trend) the overall effect on the signal to noise ratio of empirical datasets is still positive (Fig. 3E, F). That is, contra Scotland & Steel (2015) but in agreement with Wiens (2004), we do not see any evidence to support a general view that morphological dataset expansion occurs at the expense of phylogenetic signal. Increasing the number of taxa in a dataset increases homoplasy through detection of otherwise unobserved character changes. As more taxa are added to a dataset, there is an increased probability that two or more taxa have the same character state; the probability that those taxa did not derive that character through common ancestry also increases; and the likelihood of detected character loss is higher. Scotland & Steel (2015) found that datasets with

TABLE 3. Metrics of the character matrix genealogies showing the matrix-wide CI, RI and HER values, as well as the total number of characters (N char.) and parsimony-informative characters (N char. inf.).

References	CI	RI	HER	N char.	N char inf.
Early vertebrates					
Forey (1995)	0.635	0.675	0.458	56	56
Janvier (1996)	0.647	0.689	0.447	89	84
Donoghue <i>et al.</i> (2000)	0.675	0.688	0.436	103	99
Shu <i>et al.</i> (2003)	0.648	0.669	0.353	115	103
Gess <i>et al.</i> (2006)	0.672	0.678	0.416	115	103
Sansom <i>et al.</i> (2010)	0.606	0.608	0.355	109	83
Conway Morris & Caron (2014)	0.61	0.61	0.36	116	86
Gabbott <i>et al.</i> (2016)	0.651	0.674	0.457	112	79
Scalidophorans					
Wills (1998)	0.587	0.669	0.458	49	36
Dong <i>et al.</i> (2004)	0.626	0.707	0.534	87	53
Dong <i>et al.</i> (2005)	0.621	0.708	0.533	88	53
Harvey <i>et al.</i> (2010)	0.655	0.738	0.567	95	49
Wills <i>et al.</i> (2012)	0.618	0.679	0.49	88	51
Ma <i>et al.</i> (2014)	0.618	0.679	0.49	88	51
Shao <i>et al.</i> (2018)	0.608	0.671	0.475	114	48
Crocodylians					
Brochu (1997)	0.594	0.742	0.626	164	118
Brochu (1999)	0.589	0.741	0.622	164	117
Brochu (2006)	0.586	0.741	0.624	165	117
Brochu (2011)	0.582	0.739	0.621	181	118
Gold <i>et al.</i> (2014)	0.589	0.744	0.627	169	114
Lee & Yates (2018)	0.509	0.622	0.478	278	185

large proportions of homoplastic characters can lead methods, such as maximum parsimony, to produce trees that are based entirely on homoplastic characters. To address phylogenetic noise in datasets with large numbers of taxa, they used compatibility analysis to remove characters saturated with homoplasy. This improvement in accuracy comes at the expense of precision, as the removal of characters caused a general reduction in tree resolution.

Increasing the number of characters in a dataset increases the amount of homoplasy (Figs 3B, 4). However, it does not follow that adding homoplastic characters decreases phylogenetic resolution or support (Wiens 2004). This distinction is drawn out in the difference between the behaviour of the CI and RI in relation to the number of characters (Fig. 3B–D). While CI decreases with numbers of characters, RI remains approximately constant, indicating that while the consistency of additional characters may be decreasing, their phylogenetic informativeness is not; this is also reflected in

the positive trend between HER and numbers of taxa (Fig. 3E) and characters (Fig. 3F). Our analysis of matrix genealogies provides further support for this view since there is no significant decrease in CI, RI or HER through the temporal sequence of matrix development (Fig. 5), even though the dataset of Lee & Yates (2018) could represent an individual case where scoring more characters might have led to a decrease in the phylogenetic informativeness of a dataset. Furthermore, when the number of taxa and characters are considered jointly, the negative relationship between number of taxa and homoplasy (CI) remains, but number of characters is no longer significant (Table 1). Thus, the relationship between character number and homoplasy may be an artefact of the positive relationship between taxa and characters so the exhaustion of phylogenetically informative characters may not be reached in the majority of matrices. In summary, we find no support for the expectation that the phylogenetic quality of characters diminishes with discovery time (Wiens 2004).

Sanderson & Donoghue (1989) did not find a significant relationship between the number of characters and the amount of homoplasy in a dataset, though this relationship has been observed in other, smaller studies and simulations (Kluge & Farris 1969; Archie 1989; Sanderson & Donoghue 1989, 1996; Klassen *et al.* 1991; Archie & Felsenstein 1993; Lamboy 1994; Givnish & Sytsma 1997; Hauser & Boyajian 1997); we find the same results with multiple regressions despite the gradual increase in the average number of characters in morphological datasets in the intervening years: taxa, not character, number is a significant predictor of changes in homoplasy measures (Fig. 2A). Indeed, the median amount of characters in the year Sanderson & Donoghue (1989) published their study was *c.* 25, rising to a median of *c.* 150 in 2017 (Fig. 2A). This increase in the size of datasets over time is associated with increased computational power and efficiency, as well as the (not necessarily correct; Philippe *et al.* 2011) expectation that more data should yield better results (Nelson 2004). Evidently, quality, not quantity, should be the priority when choosing which taxa and characters to include in morphological datasets for phylogenetic analysis (Nelson 2004). However, an interesting corollary of this is the observation from simulated studies, that all methods of phylogenetic inference tend to support the same tree when datasets have large numbers of characters relative to numbers of taxa (O'Reilly *et al.* 2016, 2018b; Puttick *et al.* 2017b, 2019). Thus, these simulations may be producing a higher proportion of highly informative, low homoplasy characters than occurs in empirical data. However, similar trends are observed in empirical data, such as O'Leary *et al.* (2013) in which all methods tend to support a similar topology.

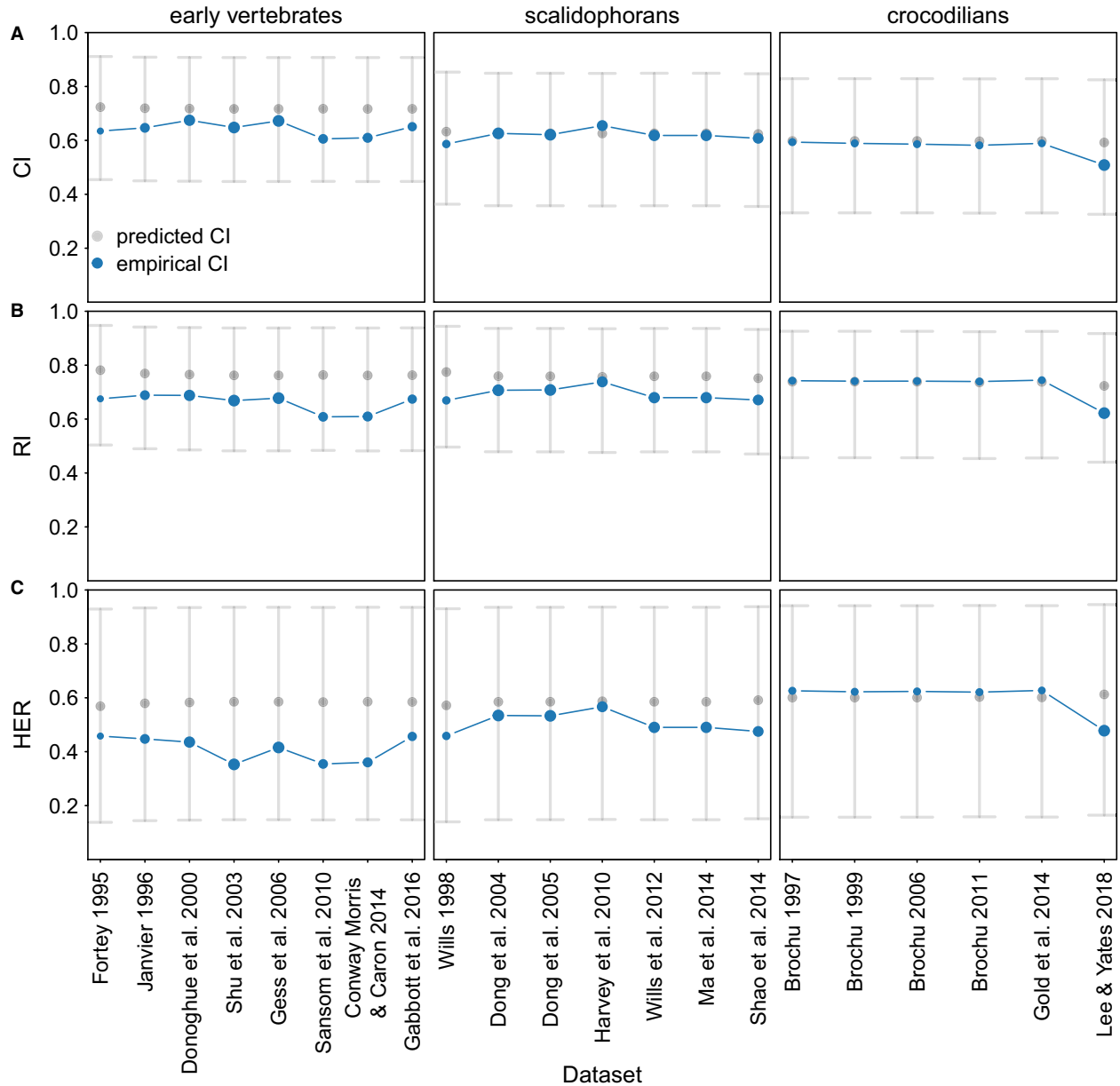


FIG. 5. Changes in homoplasy among matrix genealogies through research time. The trends in CI (A), RI (B) and HER (C) through time.

Artefacts of parsimony and homoplasy metrics

Empirical datasets have 35–40% of characters with essentially no homoplasy (i.e. CI and RI values between 0.9 and 1) (Fig. 4). This high proportion of low homoplasy characters is likely to be an artefactual consequence of these metrics being parsimony-based, measured on a most parsimonious tree. Maximum parsimony attempts to minimize the number of steps on the tree, and CI and RI are ratios of the number of steps. For this reason a high proportion of characters are expected to have the

minimum number of steps and thus have a CI or RI value of one (i.e. no homoplasy). The peak observed in the 0.4–0.5 bin in the per character CI distribution (Fig. 4) is another artefact, here reflecting the fact that 0.5 is the second highest value that a binary character can achieve after a value of one. The minimum number of steps for a binary character is always one and so a CI value of 0.5 is obtained when a binary character has just one extra step ($1/2 = 0.5$). As these parsimony-based homoplasy metrics are sensitive to such artefacts it would be desirable to develop parsimony-independent measures of homoplasy.

Outliers

A number of our analysed datasets appeared as extreme outliers on the per character CI and RI distributions, with 100% of their characters obtaining a value of one (i.e. no homoplasy) (Fig. 3A–D). The relationships among the taxa could effectively be read directly from the matrix suggesting that the range of possible characters has been filtered to exclude those with higher homoplasy. Outliers were also identified in the 0–0.1 bin on the per character RI distribution (Fig. 4), with some datasets including as much as 50% of their characters with these maximally high levels of homoplasy. Datasets composed of high proportions of highly homoplastic characters do not inspire confidence in the accuracy of the phylogenetic hypotheses derived from them.

Implications for the simulations of morphology-like data

Sanderson & Donoghue (1989) derived a formula for the expected level of CI in a morphological dataset comprised of a given number of taxa based on their survey of empirical morphological datasets. The values obtained from this survey have been used as a basis for assaying and filtering simulated morphology-like data to achieve empirical realism (O'Reilly *et al.* 2016, 2018b; Puttick *et al.* 2017b, 2019). In recent years, simulation analyses have attempted to match empirical distributions of CI to increase realism of these simulated data. Initially, this approach ensured that the overall CI profile from the simulated datasets matched the profile of overall CI exhibited by the datasets analysed by Sanderson & Donoghue (1989). Subsequently, simulations have attempted to match the average profile of CI exhibited by characters within datasets (Goloboff *et al.* 2018a; O'Reilly *et al.* 2018a; Puttick *et al.* 2019), which was not considered by Sanderson & Donoghue (1989, 1996). This approach has developed from ensuring that individual simulated datasets match the profile of CI exhibited by the universe of datasets analysed by Sanderson & Donoghue (1989), to the average profile of CI exhibited by characters within datasets (Goloboff *et al.* 2018a; O'Reilly *et al.* 2018a; Puttick *et al.* 2019), which was not considered by Sanderson & Donoghue (1989, 1996). Characterizing this phenomenon has revealed that the range of CI values exhibited by characters and datasets has increased since previous surveys (Sanderson & Donoghue 1989, 1996; Hauser & Boyajian 1997). Furthermore, the profile of CI values exhibited by characters within datasets is quite distinct from the profile of average CI exhibited by empirical datasets, as characterized by the original surveys conducted by Sanderson & Donoghue (1989, 1996). This provides a more informative target profile for future studies that attempt to simulate morphology-like data.

CONCLUSION

Our study has established a new benchmark for the distribution of CI, RI and HER in discrete character morphological datasets, revealing a greater range in the levels of matrix-wide homoplasy than has been observed in previous surveys. In particular, it seems clear that while large datasets are associated with higher levels of homoplasy, this does not necessarily imply that large datasets are less reliable or that they are inherently less phylogenetically informative. Questions have been raised about the utility of morphological datasets of ever-increasing dimensions, which may ultimately be less informative than smaller datasets, suggesting the existence of potential limits to the utility of morphology in phylogenetics. Our results show that this expectation is not met, at least as measured by the CI, RI and HER. Indeed, the approximate constancy of matrix RI relative to character number provides no evidence for the exhaustion of phylogenetically informative characters through research time, and the (slightly) positive correlation in HER between number of characters and number of taxa, indicates that there is still scope for the discovery of additional phylogenetically informative morphological characters to expand existing datasets. However, for individual studies, it might generally be worth benchmarking newly generated datasets against their progenitors to ensure that dataset expansion has not diminished phylogenetic signal. This could easily be done through comparison of RI and HER values of original and revised datasets, or by ensuring that new datasets have HER and RI values comparable to what is expected, given their size. This could be achieved by plotting the HER and RI values against those in our empirical survey (Fig. 3). The marked difference in CI and RI as characters and taxa are added to datasets, suggests that RI may be a useful characteristic for guiding the simulation of morphology-like characters for benchmarking phylogenetic methods and studying the evolution of morphology.

Acknowledgements. We thank Chris Brochu (Iowa) for access to the unpublished cladistic dataset associated with a previous publication. Peter Wagner and an anonymous referee provided constructive criticism on an earlier draft of this manuscript that led to an improved final version. This publication derives from a study undertaken by JLM in partial fulfilment of the MSc Palaeobiology at the University of Bristol in 2017–2018.

DATA ARCHIVING STATEMENT

Data for this study are available in the University of Bristol Research Data Storage Facility: <https://doi.org/10.5523/bris.sthhydmmnogk72vue56b51uftg>

Editor. Marcello Ruta

REFERENCES

- ARCHIE, J. W. 1989. Homoplasy excess ratios: new indices for measuring levels of homoplasy in phylogenetic systematics and a critique of the consistency index. *Systematic Zoology*, **38**, 235–269.
- 1996. Measures of homoplasy. 153–188. In SANDERSON, M. J. and HUFFORD, L. (eds). *Homoplasy: The recurrence of similarity in evolution*. Academic Press.
- and FELSENSTEIN, J. 1993. The number of evolutionary steps on random and minimum length trees for random evolutionary data. *Theoretical Population Biology*, **43**, 52–79.
- BROCHU, C. A. 1997. Morphology, fossils, divergence timing, and the phylogenetic relationships of *Gavialis*. *Systematic Biology*, **46**, 479–522.
- 1999. Taxon sampling and reverse successive weighting. *Systematic Biology*, **48**, 808–813.
- 2006. Osteology and phylogenetic significance of *Eosuchus minor* (Marsh, 1870) new combination, a longirostrine crocodylian from the Late Paleocene of North America. *Journal of Paleontology*, **80**, 162–186.
- 2011. Phylogenetic relationships of *Necrosuchus ionensis* Simpson, 1937 and the early history of caimanines. *Zoological Journal of the Linnean Society*, **163**, S228–S256.
- BROWN, J. W., PARINS-FUKUCHI, C., STULL, G. W., VARGAS, O. M. and SMITH, S. A. 2017. Bayesian and likelihood phylogenetic reconstructions of morphological traits are not discordant when taking uncertainty into consideration: a comment on Puttick et al. *Proceedings of the Royal Society B*, **284**, 20170986.
- CONWAY MORRIS, S. C. and CARON, J. B. 2014. A primitive fish from the Cambrian of North America. *Nature*, **512**, 419–422.
- CRIBARI-NETO, F. and ZEILEIS, A. 2010. Beta regression in R. *Journal of Statistical Software*, **34**, 1–24.
- DONG, X.-P., DONOGHUE, P. C. J., CHENG, H. and LIU, J. 2004. Fossil embryos from the Middle and Late Cambrian period of Hunan, south China. *Nature*, **427**, 237–240.
- CUNNINGHAM, J., LIU, J. and CHENG, H. 2005. The anatomy, affinity and phylogenetic significance of *Markuelia*. *Evolution & Development*, **7**, 468–482.
- DONOGHUE, P. C. J., FOREY, P. L. and ALDRIDGE, R. J. 2000. Conodont affinity and chordate phylogeny. *Biological Reviews*, **75**, 191–251.
- FARRIS, J. S. 1989. The retention index and the rescaled consistency index. *Cladistics*, **5**, 417–419.
- 1991. Excess homoplasy ratios. *Cladistics*, **7**, 81–91.
- FOREY, P. L. 1995. Agnathans recent and fossil, and the origin of jawed vertebrates. *Reviews in Fish Biology & Fisheries*, **5**, 267–303.
- GABBOTT, S. E., DONOGHUE, P. C., SANSOM, R. S., VINTHER, J., DOLOCAN, A. and PURNELL, M. A. 2016. Pigmented anatomy in carboniferous cyclostomes and the evolution of the vertebrate eye. *Proceedings of the Royal Society B*, **283**, 20161151.
- GESS, R. W., COATES, M. I. and RUBIDGE, B. S. 2006. A lamprey from the Devonian period of South Africa. *Nature*, **443**, 981–984.
- GIVNISH, T. J. and SYTSMA, K. J. 1997. Consistency, characters, and the likelihood of correct phylogenetic influence. *Molecular Phylogenetics & Evolution*, **7**, 320–330.
- GOLD, M. E. L., BROCHU, C. A. and NORELL, M. A. 2014. An expanded combined evidence approach to the *Gavialis* problem using geometric morphometric data from crocodylian braincases and eustachian systems. *PLoS One*, **9**, e105793.
- GOLOBOFF, P. A., TORRES, A. and ARIAS, J. S. 2018a. Parsimony and model-based phylogenetic methods for morphological data: a comment on O'Reilly et al. (2017). *Palaeontology*, **61**, 625–630.
- 2018b. Weighted parsimony outperforms other methods of phylogenetic inference under models appropriate for morphology. *Cladistics*, **34**, 407–437.
- HARVEY, T. H. P., DONG, X. and DONOGHUE, P. C. J. 2010. Are palaeoscoleids ancestral ecdysozoans? *Evolution & Development*, **12**, 177–200.
- HAUSER, D. L. and BOYAJIAN, G. 1997. Proportional change and patterns of homoplasy: Sanderson and Donoghue revisited. *Cladistics*, **13**, 97–100.
- JAMES, G., WITTEN, D., HASTIE, T. and TIBSHIRANI, R. 2013. *An introduction to statistical learning with applications in R*. Springer, 426 pp.
- JANVIER, P. 1996. The dawn of the vertebrates: characters versus common ascent in the rise of current vertebrate phylogenies. *Palaeontology*, **39**, 259–287.
- KEATING, J. N., SANSOM, R. S., SUTTON, M. D., KNIGHT, C. G. and GARWOOD, R. J. 2020. Morphological phylogenetics evaluated using novel evolutionary simulations. *Systematic Biology*, **69**, 897–912.
- KLASSEN, G. J., MOOI, R. D. and LOCKE, A. 1991. Consistency indices and random data. *Systematic Zoology*, **40**, 446–457.
- KLUGE, A. G. and FARRIS, J. S. 1969. Quantitative phyletics and the evolution of anurans. *Systematic Biology*, **18**, 1–32.
- LAMBOY, W. F. 1994. The accuracy of the maximum parsimony method for phylogeny reconstruction with morphological characters. *Systematic Botany*, **19**, 489–505.
- LEE, M. S. Y. and YATES, A. M. 2018. Tip-dating and homoplasy: reconciling the shallow molecular divergences of modern gharials with their long fossil record. *Proceedings of the Royal Society B*, **285**, 20181071.
- MA, X., ALDRIDGE, R. J., SIVETER, D. J., SIVETER, D. J., HOU, X. and EDGEcombe, G. D. 2014. A new exceptionally preserved Cambrian priapulid from the Chengjiang Lagerstätte. *Journal of Paleontology*, **88**, 371–384.
- MURPHY, J. L., PUTTICK, M. N., O'REILLY, J. E., PISANI, D. and DONOGHUE, P. C. J. 2021. Data from Murphy et al. 2021. Empirical distributions of homoplasy in morphological data. Palaeontology. University of Bristol Research Data Storage Facility (data.bris). <https://doi.org/10.5523/bris.sthhydmgk72vue56b51uftg>
- NELSON, G. 2004. Cladistics: its arrested development. 127–147. In WILLIAMS, D. M. and FOREY, P. L. (eds). *Milestones in systematics*. Systematics Association Special Volume **47**, CRC Press.
- O'LEARY, M. A., BLOCH, J. I., FLYNN, J. J., GAUDIN, T. J., GIALLOMBARDO, A., GIANNINI, N. P., GOLDBERG, S. L., KRAATZ, B. P., LUO, Z.-X., MENG, J.,

- NI, X., NOVACEK, M. J., PERINI, F. A., RANDALL, Z. S., ROUGIER, G. W., SARGIS, E. J., SILCOX, M. T., SIMMONS, N. B., SPAULDING, M., VELAZCO, P. M., WEKSLER, M., WIBLE, J. R. and CIRRANELLO, A. L. 2013. The placental mammal ancestor and the post-K-Pg radiation of placentals. *Science*, **339**, 662–667.
- O'REILLY, J. E., PUTTICK, M. N., PARRY, L. A., TANNER, A. R., TARVER, J. E., FLEMING, J., PISANI, D. and DONOGHUE, P. C. J. 2016. Bayesian methods outperform parsimony but at the expense of precision in the estimation of phylogeny from discrete morphological data. *Biology Letters*, **12**, 20160081.
- PISANI, D. and DONOGHUE, P. C. J. 2018a. Empirical realism of simulated data is more important than the model used to generate it. A reply to Goloboff et al. *Palaeontology*, **61**, 631–635.
- ——— 2018b. Probabilistic methods surpass parsimony when assessing clade support in phylogenetic analyses of discrete morphological data. *Palaeontology* **61**, 105–118.
- PHILIPPE, H., BRINKMANN, H., LAVROV, D. V., LITTLEWOOD, D. T. J., MANUEL, M., WÖRHEIDE, G. and BAURAIN, D. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biology*, **9**, e1000602.
- PUTTICK, M. N., O'REILLY, J. E., OAKLEY, D., TANNER, A. R., FLEMING, J. F., CLARK, J., HOLLOWAY, L., LOZANO-FERNANDEZ, J., PARRY, L. A., TARVER, J. E., PISANI, D. and DONOGHUE, P. C. J. 2017a. Parsimony and maximum-likelihood phylogenetic analyses of morphology do not generally integrate uncertainty in inferring evolutionary history: a response to Brown et al. *Proceedings of the Royal Society B*, **284**, 20171636.
- TANNER, A. R., FLEMING, J. F., CLARK, J., HOLLOWAY, L., LOZANO-FERNANDEZ, J., PARRY, L. A., TARVER, J. E., PISANI, D. and DONOGHUE, P. C. J. 2017b. Uncertain-tree: discriminating among competing approaches to the phylogenetic analysis of phenotype data. *Proceedings of the Royal Society B*, **284**, 20162290.
- PISANI, D., DONOGHUE, P. C. J. and RAHMAN, I. 2019. Probabilistic methods outperform parsimony in the phylogenetic analysis of data simulated without a probabilistic model. *Palaeontology*, **62**, 1–17.
- R CORE TEAM 2016. *R: a language and environment for statistical computing*. v. 3.3.2. R Foundation for Statistical Computing. <https://www.R-project.org>
- SANDERSON, M. J. and DONOGHUE, M. J. 1989. Patterns of variation in levels of homoplasy. *Evolution*, **43**, 1781–1795.
- 1996. The relationship between homoplasy and confidence in a phylogenetic tree. 67–89. In SANDERSON, M. J. and HUFFORD, L. (eds). *Homoplasy: The recurrence of similarity in evolution*. Academic Press.
- SANSOM, R. S., CHOATE, P. G., KEATING, J. N. and RANDLE, E. 2018. Parsimony, not Bayesian analysis, recovers more stratigraphically congruent phylogenetic trees. *Biology Letters*, **14**.
- FREEDMAN, K., GABBOTT, S. E., ALDRIDGE, R. J. and PURNELL, M. A. 2010. Taphonomy and affinity of an enigmatic Silurian vertebrate, *Jamoytius kerwoodi* White. *Palaeontology*, **53**, 1393–1409.
- and WILLS, M. A. 2017. Differences between hard and soft phylogenetic data. *Proceedings of the Royal Society B*, **284**, 20172150.
- SCOTLAND, R. W., OLMSTEAD, R. G. and BENNETT, J. R. 2003. Phylogeny reconstruction: the role of morphology. *Systematic Biology*, **52**, 539–548.
- and STEEL, M. 2015. Circumstances in which parsimony but not compatibility will be provably misleading. *Systematic Biology*, **64**, 492–504.
- SHAO, T., TANG, H., LIU, Y., WALOSZEK, D., MAAS, A. and ZHANG, H. 2018. Diversity of cnidarians and cycloneuralians in the Fortunian (early Cambrian) Kuan-chuanpu Formation at Zhangjiagou, South China. *Journal of Paleontology*, **92**, 115–129.
- SHU, D.-G., CONWAY MORRIS, S., HAN, J., ZHANG, Z.-F., YASUI, K., JANVIER, P., CHEN, L., ZHANG, X.-L., LIU, J.-N. and LIU, H.-Q. 2003. Head and backbone of the Early Cambrian vertebrate *Haikouichthys*. *Nature*, **421**, 526–529.
- SWOFFORD, D. L. 2002. PAUP*: Phylogenetic Analysis Using Parsimony (*and other methods). Sinauer Associates, Sunderland, MA.
- WIENS, J. 2004. The role of morphological data in phylogeny reconstruction. *Systematic Biology*, **53**, 653–661.
- WILLS, M. A. 1998. Cambrian and recent disparity: the picture from priapulids. *Paleobiology*, **24**, 177–199.
- GERBER, S., RUTA, M. and HUGHES, M. 2012. The disparity of priapulid, archaeopriapulid and palaeoscolecoid worms in the light of new data. *Journal of Evolutionary Biology*, **25**, 2056–2076.
- WRIGHT, A. M., LLOYD, G. T. and HILLIS, D. M. 2016. Modeling character change heterogeneity in phylogenetic analyses of morphology through the use of priors. *Systematic Biology*, **65**, 602–611.